Working Paper 93-27
Statistics and Econometrics Series 21
October, 1993

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

COMPUTING MISSING VALUES IN TIME SERIES

Victor Gomez, Agustín Maravall and Daniel Peña*

Abstract

This work presents two algorithms to estimate missing values in time series. The first is the Kalman Filter, as developed by Kohn and Ansley (1986) and others. The second is the additive outlier approach, developed by Peña, Ljung and Maravall. Both are exact and lead to the same results. However, the first is, in general, faster and the second more flexible.

Key Words

Kalman filtering, additive outliers, nonstationary ARIMA processes, concentrated likelihoods.

*Gómez, Instituto Nacional de Estadística, 28046 Madrid (Spain); Maravall, European University Institute, I-50016 S, Domenico di Fiesole (FI), Italy; Peña, Departamento de Estadística y Econometría. Universidad Carlos III de Madrid.

Computing Missing Values in Time Series

Victor Gómez, Agustín Maravall and Daniel Peña

Instituto Nacional de Estadística. 28046 Madrid, Spain.

European University Institute. I-50016 S. Domenico di Fiesole (FI), Italy.

Universidad Carlos III de Madrid. c/ Madrid, 126. 28903 Getafe, Madrid. Spain.

Abstract

This work presents two algorithms to estimate missing values in time series. The first is the Kalman Filter, as developed by Kohn and Ansley (1986) and others. The second is the additive outlier approach, developed by Peña, Ljung and Maravall. Both are exact and lead to the same results. However, the first is, in general, faster and the second more flexible.

Key words: Kalman Filtering, additive outliers, nonstationary ARIMA processes, concentrated likelihoods.

1. INTRODUCTION

The analysis of time series represented by ARIMA models when some data points are missing has received considerable attention in the literature. Brubacher and Wilson (1976) developed by least squares an interpolation procedure that led to an estimator which is a linear function of the known terms in the series and has minimum squared error. Miller and Ferreiro (1984) showed that the least squares estimators of the missing values are equivalent to the conditional expectations of the missing observations given the data and the parameters of the model. This result can also be obtained directly from the decomposition of the exact likelihood function of an ARMA process with missing data made by Ljung (1982). Jones (1980) used the state space representation of an ARIMA model and the Kalman Filter to compute the likelihood of an ARMA model, and showed how to use this recursive estimation procedure to estimate the parameters of the model when some observations are missing. Then, in order to estimate the missing values the fixed point smoother can be used. This approach was extended by Ansley and Kohn (1983), Harvey and Pierse (1984) and Kohn and Ansley (1983, 1986), to the nonstationary case. The numerical problems involved in the maximization of the likelihood were analyzed by Wincek and Reinsel (1986). Kohn and Ansley (1986) introduced a general definition of the likelihood of a non-stationary ARIMA model that allowed, for the first time, the incorporation of missing values in the pre-observation period of the series required to initialize the computations. The approach of these last authors resolved the problem, but required a modified Kalman filter to compute the likelihood and to predict future observations. Bell and Hillmer (1991) have shown that the

same results could be obtained with a suitable initialization of the ordinary Kalman filter. Gómez and Maravall (1992) have presented an alternative definition of the likelihood that can be used with the standard Kalman Filter and, thus, does not require any modification of existing computational routines.

Peña (1987) showed the relationship between missing value interpolation, additive outlier estimation, inverse autocorrelations and measures of data influence in time series models. The relationship between missing values and additive outliers has also been explored by Ljung (1989). Pourahmadi (1989) presented the estimation and interpolation problem from the point of view of the EM algorithm. Peña and Maravall (1991) analyzed the general case of any possible distribution of missing observations in an ARIMA time series model, with known model parameters and obtained analytical expressions for the optimal estimators and their associated mean squared errors, that involve solely the elements of the inverse autocorrelation function of the series. This approach leads to a different estimation procedure for the missing values based on replacing the missing values in the series with arbitrary numbers and treating then these numbers as additive outliers. This method, that will be called the Peña-Ljung-Maravall procedure in this paper, leads to an efficient algorithm for parameter estimation and interpolation when the number of missing data is moderate.

This paper analyzes these two main procedures for estimating missing data in time series and compares them from a computational point of view. The work is organized as follows. Section 2 presents the likelihood function of a non-stationary ARIMA process with missing data. Section 3 reviews the recursive approach to interpolation using the Kalman filter. Section 4 analyzes the Peña-Ljung-Maravall method based on additive outlier estimation of the missing values. Section 5 presents the computational performance of these procedures.

2. THE LIKELIHOOD OF AN ARIMA MODEL WITH MISSING DATA

2.1 The Stationary case

Suppose that over a time sequence of n periods we observe the discrete time series

$$\boldsymbol{Z} = (Z_{t_1}, \ldots, Z_{t_m}), \qquad t_1 \leq t_2 \leq \ldots \leq t_m$$

where m<n and, therefore, we will have n-m=h missing data points. We assume that the complete data set $\tilde{Z}=(Z_1,\ldots,Z_n)$, that includes $\mathbf{Z}$ and the vector $\mathbf{Z}_a$ of h missing values, follows the univariate ARMA (p,q) model

$$\phi(B)\, Z_t = \theta(B)\, a_t \tag{2.1}$$

where $\phi(B) = (1-\phi_1 B - \ldots - \phi_p B^p)$ and $\theta(B)=(1-\theta_1 B-\ldots-\theta_q B^q)$ have the roots outside the unit circle and do not have common roots, B is the backshift operator, and a_t is an i.i.d. $N(0,\sigma^2)$ process. Let us assume that the observed vector of data $\mathbf{Z}$ has zero expected values and $\sigma^2\Omega$ covariance matrix, then, calling $\boldsymbol{\beta}$ the vector of parameters of model (2.1) the likelihood function is

$$l(\beta \mid Z) = \frac{1}{(2\pi\sigma^2)^{m/2}\,|\Omega|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{Z}'\Omega^{-1}\boldsymbol{Z}\right\}. \tag{2.2}$$

Of course, when m=n (2.2) is the usual likelihood function of a stationary ARMA model.

2.2 The nonstationary case

Let us consider the case of a non-stationary time series that follows the ARIMA (p,d,q) model

$$\phi(B)\,\delta(B)\,Z_t=\theta(B)\,a_t \tag{2.3}$$

where $\phi(B)$, $\theta(B)$ and a_t are the same as in (2.1), and $\delta(B) = (1-\delta_1 B-\ldots-\delta_d B^d)$ has all its roots on the unit circle, and includes the differencing operators. Calling

$$u_t=\delta(B)\,Z_t \qquad t=d+1,\ldots,n \tag{2.4}$$

the values of the stationary transformation, the likelihood function for a non-stationary ARIMA process without missing values is defined by using the marginal density of u_t. Therefore

$$l(\beta\,|\,\boldsymbol{u}) = \frac{1}{(2\pi\sigma^2)^{\frac{n-d}{2}}\,|\Omega_u|^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}\,\boldsymbol{u}'\Omega_u^{-1}\boldsymbol{u}\right\} \tag{2.5}$$

where $\sigma^2\Omega_u$ is the covariance matrix of the normal vector **u**, that has n-d components.

Differencing a time series with missing data will introduce additional missing values into the differences series. Consequently, the likelihood function in this case needs to be written as a function of the original data **Z**. In order to express (2.5) in this way, let $\mathbf{Z}'=(Z'_*, Z'_R)'$ be the sample data, where Z'_* is the $(1\times d)$ vector of starting values for (2.4) and Z'_R the $1\times(n-d)$ vector of remaining observations. Then (2.4) implies the transformation

$$\begin{bmatrix} Z_* \\ u \end{bmatrix} = \begin{bmatrix} I_d & 0 \\ T_1 & T_2 \end{bmatrix}\begin{bmatrix} Z_* \\ Z_R \end{bmatrix} \tag{2.6}$$

where I_d is the identity matrix of rank d and T_1 and T_2 are triangular matrices given by

$$T_1 = \begin{bmatrix} -\delta_d & \cdots & -\delta_1 \\ & \ddots & \\ & & \ddots \\ & & & -\delta_d \\ 0 & & & \end{bmatrix},\quad T_2 = \begin{bmatrix} 1 & & & & \\ -\delta_1 & 1 & & & \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ -\delta_d & & & & \ddots \\ & -\delta_d & -\delta_1 & & 1 \end{bmatrix} \tag{2.7}$$

The data **Z** is related to the starting values Z_* and to the stationary variable u by

$$\boldsymbol{Z} = \begin{bmatrix} Z_* \\ Z_R \end{bmatrix} = \begin{bmatrix} I_d & 0 \\ -T_2^{-1}T_1 & T_2^{-1} \end{bmatrix} \begin{bmatrix} Z_* \\ u \end{bmatrix} \tag{2.8}$$

and calling $A=-T_2^{-1}T_1$ and

$$v = T_2^{-1}u, \tag{2.9}$$

where A and v can be computed recursively as shown in Bell (1984), we have

$$Z_R=AZ_*+v \tag{2.10}$$

In order to write the density function for **Z** we need some assumptions on the marginal distribution of Z_*. Assuming that it is normal, the joint distribution of (Z_*, Z_R) will be multivariate normal, as is the one of (Z_*, u), and the Jacobian of the transformation is, according to (2.6) equal to unity. Therefore

$$f(Z_R|Z_*) = f(u|Z_*) \tag{2.11}$$

Making the assumption that **u** is independent of Z_*, the conditional distribution f(u | Z_*) is identical to the marginal, which leads to likelihood (2.5). Also, as Z_R | Z_* is, for (2.10), normal with mean AZ_*, and calling Σ the covariance matrix of v, the likelihood (2.5) can also be written as

$$l(\beta|Z_R) = \frac{1}{(2\pi\sigma^2)^{\frac{n-d}{2}}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(Z_R-AZ_*)'\boldsymbol{\Sigma}^{-1}(Z_R-AZ_*)\right\} \tag{2.12}$$

Equation (2.12) suggests an alternative definition of the likelihood function when some observations are missing. Supose that we have observed a sample of size m of nonconsecutive observations of a time series that follows (2.3), and let us assume that the vector Z_* of starting values is $\{Z_I, Z_J\}$, where Z_I is the vector of observed data, that we assumed has k components ($k \leq d$), and Z_J the vector of missing data with d-k components, and Z_R is $\{Z_o, Z_a\}$, where Z_o includes the observed data, that we assumed has m-k components, and Z_a the missing data (with n-m-(d-k) components). Then (2.10) can be written as

$$Z_R=BZ_I+CZ_J+v \tag{2.13}$$

where B and C include the rows of A corresponding to the observed and missing data and, consequently, B is (n-d)×k and C is (n-d)×(d-k). Then, we can also partition Z_R into the observed and missing parts and obtain for the observed data

$$Z_o=B_oZ_I+C_oZ_J+v_o \tag{2.14}$$

where now B_o is (m-k)×k, C_o is (m-k)×(d-k), and v_o is a (m-k)×1 vector. Then, it is reasonable to define the likelihood function using the distribution of Z_o conditional on Z_I and

Z_J, as before. The main difference from the stationary case is that now Z_J will be an unknown parameter to be estimated. Calling

$$y_o = Z_o - B_o Z_I \tag{2.15}$$

the likelihood function (2.12) will be written in this case as

$$l(\beta, Z_J | Z_o) = \frac{1}{(2\pi\sigma^2)^{\frac{m-k}{2}} |\Sigma_o|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(y_o - C_o Z_J)' \Sigma_o^{-1} (y_o - C_o Z_J)\right\} \tag{2.16}$$

where y_o is known Σ_o is the covariance matrix of v_o and Z_J is a vector of parameters to be estimated. This approach has been proposed by Gómez and Maravall (1992) as they showed this definition of the likelihood is equivalent to the one suggested by Kohn and Ansley (1986). Note that (2.16) is also the likelihood of the regression model

$$y_o = C_o Z_J + v_o \tag{2.17}$$

result that will be used in the next sections.

3. COMPUTATIONS USING THE KALMAN FILTER

3.1 The Stationary case

A stationary time series that follows an ARMA model can be represented by the general AR(∞) model (Box and Jenkins, 1976)

$$\pi(B) Z_t = a_t \tag{3.1}$$

where $\pi(B) = \phi(B)\theta(B)^{-1}$. Assuming that the sample size is large, and that the unobserved a_t for $t<0$ are zero, (3.1) can be written in matrix form using the approximation

$$\pi \mathbf{Z} = \mathbf{e} \tag{3.2}$$

where π is a lower triangular matrix with ones in the main diagonal and coefficients $\{-\pi_i\}$ in the rows. Thus, calling $\Omega_Z\sigma^2$ the covariance matrix of the vector $\mathbf{Z}$ we have

$$\mathbf{Z} = \pi^{-1}\mathbf{e} \tag{3.3}$$

and

$$\Omega_Z = (\pi^{-1})(\pi^{-1})'. \tag{3.4}$$

Expression (3.4) shows that the computation of the residual e can be done by using the Cholesky decomposition of the covariance matrix of the process. The Kalman filter can be seen as an exact and efficient recursive algorithm to obtain this Cholesky decomposition. This algorithm computes the vector of one step ahead residuals $\hat{a}_i$, and its variances, given

by $l_{ii}\sigma^2$, where the elements l_{ii} are the diagonal elements of the decomposition

$$\Omega = LL'. \tag{3.5}$$

The likelihood function (2.2) can be written as a function of these statistics as:

$$L(\theta|Z) = \log l(\theta|Z) = -\frac{m}{2} \log \sigma^2 - \Sigma \log l_{ii} - \frac{1}{2\sigma^2} \hat{a}'\hat{a} \tag{3.6}$$

and, for given values of the parameters ϕ and θ, the maximum of this function with respect to σ^2 is always attained at $\sigma^2 = \hat{a}'\hat{a}/m$. Therefore, we can concentrate σ^2 out of (3.6) and maximize:

$$L^*(\theta|Z) = constant - \frac{m}{2} \log \hat{a}'\hat{a} - \Sigma \log l_{ii} = constant - \frac{m}{2} \log S \tag{3.7}$$

where

$$S = |L|^{1/m} \hat{a}'\hat{a} |L|^{1/m} \tag{3.8}$$

This procedure is general and can be applied with and without missing values. In the first case, the terms l_{ii} are computed directly by the Kalman Filter, as shown in Jones (1980), and the minimization of (3.8) will provide an estimation of the parameters in the case of missing values. The estimation of the missing observations can then be computed by using the fixed point smoother algorithm (see Anderson and Moore, 1979).

3.2 The non-stationary case

For nonstationary series using (2.16) and (2.17) the Kalman filter will provide the residuals $L_0^{-1}(y_o - C_o Z_J)$ and the diagonal element l_{ii} needed to compute $|L_o|$, where now $\Sigma_o = L_o L_o'$. Then, the new vector of parameters (ϕ, θ, Z_J) can be obtained by minimizing

$$S = |L_o|^{\frac{1}{m-k}} (L_o^{-1}y_o - L_o^{-1}C_o Z_J)'(L_o^{-1}y_o - L_o^{-1}C_o Z_J) |L_o|^{\frac{1}{m-k}} \tag{3.9}$$

and

$$\hat{\sigma}^2 = \frac{1}{m-k} (L_o^{-1}y_o - L_o^{-1}C_o Z_J)'(L_o^{-1}y_o - L_o^{-1}C_o Z_J) . \tag{3.10}$$

A more efficient procedure can be obtained by concentrating Z_J out of (3.9). In order to do so, the Kalman filter is applied to both the vector y_o, and to the columns of the matrix C_o to obtain

$$L_o^{-1}y_o = L_o^{-1}C_o Z_J + L_o^{-1}v_o, \tag{3.11}$$

now, the QR algorithm applied to $L_o^{-1}C_o$ provides an orthogonal matrix $Q' = (Q_1\ Q_2)'$ which verifies

$$Q_1' L_o^{-1} y_o = R Z_J + Q_1' L_o^{-1} v_o \tag{3.12}$$

$$Q_2' L_o^{-1} y_o = \qquad Q_2' L_o^{-1} v_o \tag{3.13}$$

and, therefore Z_J is estimated by

$$\hat{Z}_J = R^{-1} Q_1' L_o^{-1} y_o \tag{3.14}$$

and S can be written as

$$S = |L_o|^{\frac{1}{m-k}} (Q_2' L_o^{-1} y_o)' (Q_2' L_o^{-1} y_o) |L_o|^{\frac{1}{m-k}}. \tag{3.15}$$

The parameters (ϕ, θ) are estimated by minimizing (3.15) and the interpolated values are obtained by a smoothing algorithm. See Gómez and Maravall (1992) for the efficient use of the Kalman Filter to carry out the computations.

4. COMPUTATIONS USING THE PEÑA-LJUNG-MARAVALL PROCEDURE

4.1 The stationary case

An alternative approach for the estimation of the missing values is to fill the gaps in the series with arbitrary numbers, and estimate the parameters and the missing data by using the relationship between additive outlier estimation and optimal interpolation indicated in Peña (1987), Ljung (1989) and Peña and Maravall (1991). Starting with the stationary case, let us call, as before, $\tilde{\mathbf{Z}}$ the complete vector and $\mathbf{Z}_a$ and $\mathbf{Z}$ the missing and observed vectors. Then, the following relation among the densities:

$$f(\mathbf{Z}) = \frac{f(\tilde{\mathbf{Z}})}{f(\mathbf{Z}_a/\mathbf{Z})} \tag{4.1}$$

implies that, calling Ω, $\tilde{\Omega}$ and Ω_a the covariance matrices for the distribution of $\mathbf{Z}$, $\tilde{\mathbf{Z}}$ and Z_a/Z respectively,

$$|\Omega| = |\tilde{\Omega}| |\Omega_a|^{-1} \tag{4.2}$$

$$Z'\Omega^{-1}Z = \tilde{Z}'\tilde{\Omega}^{-1}\tilde{Z} - (Z_a - PZ)'\Omega_a^{-1}(Z_a - PZ) \tag{4.3}$$

where $E(Z_a \mid Z) = PZ$ is the minimum mean squared error interpolator of the vector of the missing values. Now, let Z_C be the series completed by filling the missing values with arbitrary numbers Z_a. Calling

$$w = \overline{Z}_a - Z_a \tag{4.4}$$

we can write the vector $\tilde{Z}$ of the complete unobserved series as

$$\tilde{Z}=Z_C-Xw \tag{4.5}$$

where X is a (n×h) matrix such that its columns are dummy variables (that is, there is a value equal to one and zero otherwise) corresponding to the h missing data. Then, by (4.5) we have transformed the unobserved $\tilde{Z}$ into a completely known series Z_C but with h additive outliers w. The optimal estimate of w can be obtain by inserting (4.5) and (4.4) in (4.3), with the following result:

$$(Z_C-Xw)'\Omega^{-1}(Z_C-Xw)=Z'\Omega^{-1}Z+(\overline{Z}_a-w-PZ)'\Omega_a^{-1}(Z_a-w-PZ)\ . \tag{4.6}$$

The estimation of w requires the minimization of the right-hand side. This is clearly achieved by setting:

$$\hat{w}=\overline{Z}_a-PZ=\overline{Z}_a-E(Z_a|Z) \tag{4.7}$$

that means that the estimator of w is the difference between the arbitrary inserted value and the optimal interpolator, computed by the expected value of the missing values given the rest of the data. This estimator can also be interpreted seeing that, for fixed Z_a, w is a random variable according to (4.4) with a normal distribution. The minimum square error estimator of w will be its mean, and taking expectations conditional to the observed data Z in (4.4) again result (4.7) is obtained. On the other hand, the value that minimize the left-hand side is the generalized least square estimator

$$\hat{w}=(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Z_C \tag{4.8}$$

with estimated covariance matrix

$$v(\hat{w})=\sigma^2(X'\Omega^{-1}X)^{-1} \tag{4.9}$$

Therefore, as (4.7) and (4.8) minimize (4.6) both must be the same. Also, as for (4.4) and (4.7)

$$w-\hat{w}=-(Z_a-E(Z_a/Z))$$

$v(\hat{w})$ is also equal to Ω_a. It is interesting to note that in (4.4) w is treated as a random variable, whereas in (4.8) it is treated as a parameter. A discussion of the conditions which leads to the same estimate in these cases can be found in Peña and Tiao (1991).

To write the likelihood function for β given Z, we first note that $|\Omega|$ can be written for (4.2) and the expression (4.9) for Ω_a as a function of the $\tilde{\Omega}$ matrix. Also, the exponent can be written as a function of $\tilde{\Omega}$ by using (4.7) in (4.6):

$$Z'\Omega^{-1}Z=(Z_C-X\hat{w})'\Omega^{-1}(Z_C-X\hat{w}) \tag{4.11}$$

and, therefore, the likelihood function for β given Z is

$$l(\beta|Z)=(2\pi\sigma^2)^{-n/2}|\Omega|^{-1/2}|X'\Omega^{-1}X|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(Z_C-X\hat{w})'\Omega^{-1}(Z_C-X\hat{w})\right\} \tag{4.12}$$

The maximization of (4.12) to obtain the parameters can be carried out again using

the Kalman Filter, as shown in section 3.2, or in the following steps. (Step 1) insert arbitrary values in the missing values position (for instance, the mean of the series or the mean of the two adjacent points); (Step 2) estimate the parameters with the complete data set by the usual procedure; (Step 3) assume additive outliers at the missing value positions and estimate their magnitude by (4.8); (Step 4) correct the series of outliers by Z_C - X $\hat{w}$ and estimate again the parameters by the usual procedure *but* inserting an additional term $| X' \tilde{\Omega}^{-1} X |^{-1/2}$ in the likelihood. Repeat Step 3 and Step 4 until convergence. This procedure was first suggested by Peña (1987) for the case of a single missing point and extended by Ljung (1989), and Peña and Maravall (1991).

It is worth stressing that if we have a complete series and assume that a vector of h additive outliers is affecting it, the likelihood function differs from (4.12) only by the determinant $| X'\tilde{\Omega}X |^{-1/2}$. This fact suggests that an approximate procedure to estimate missing values in time series is to introduce arbitrary values at the missing positions and then use any standard routine that allows for the inclusion of dummy variables in a time series, as discussed by Box and Tiao (1975). This intervention analysis approach can be easily carried out by standard software.

4.2 The non-stationary case

For non-stationary series it is convenient to express the likelihood as a function of the original data. Let $\tilde{Z}_R$ be, as previously, the complete unobserved series and let Z_o and Z_a be the set of observed and missing values for t>d. Then letting $y_o=C_oZ_J+v_o=Z_o-B_oZ_I$ as defined in (2.15), and $\tilde{y}=\tilde{Z}_R-BZ_I$, $y_a=Z_a-B_aZ_I$, conditioning on Z_J, $\tilde{y}$, y_o and y_a are normal random variables, and we can write

$$f(y_o) = \frac{f(\tilde{y})}{f(y_a|y_o)} \tag{4.13}$$

and $f(y_o)$ leads to the likelihood (2.16). We want, as in the stationary case, to express this likelihood as a function of $\tilde{\Omega}$, the standard covariance matrix. To achive this objective, the same procedure used to obtain the formulas from (4.2) to (4.11) can be applied by using (y_o, $\tilde{y}$, y) instead of (Z_o, $\tilde{Z}$, Z) and (Σ_o, $\tilde{\Sigma}$, Σ_a) instead of (Ω, $\tilde{\Omega}$, Ω_a). Note that Σ_o, $\tilde{\Sigma}$, Σ_a are the covariance matrices for (y_o, $\tilde{y}$, y_a) and, also, the covariance matrices of the vectors (v_o, $\tilde{v}$, v_a). Therefore, (4.13) implies

$$|\Sigma_o| = |\tilde{\Sigma}| \, |\Sigma|^{-1} \tag{4.14}$$

and

$$v_o'\Sigma_o^{-1}v_o = \tilde{v}'\tilde{\Sigma}^{-1}\tilde{v} - (v_a-P_av_o)'\Sigma_a^{-1}(v_a-P_av_o) \tag{4.15}$$

where $E(y_a \mid y_o) = C_aZ_J + P_av_o$. Now, let $\bar{Z}_a$ be the vector of arbitrary values inserted at the missing value positions, and let

$$w_a=\bar{Z}_a-Z_a$$

where we use w_a instead of w to differentiate the stationary and non stationary case. Defining, as before, the matrix X_a of dummy variables

$$\tilde{Z}_R = Z_C - X_a w_a,$$

where Z_C is the series completed with $\bar{Z}_a$, and subtracting $AZ_\bullet = BZ_I + CZ_J$ from both sides

$$\tilde{v} = (\tilde{y} - CZ_J) = (y_C - CZ_J) - X_a w_a = v_C - X_a w_a \tag{4.16}$$

where $y_c = Z_C - BZ_I$. It can be shown by introducing (4.16) into (4.15), that the optimun estimator of w_a is

$$\hat{w}_a = (X_a' \tilde{\Sigma}^{-1} X_a)^{-1} X_a' \tilde{\Sigma}^{-1} (y_C - CZ_J) \tag{4.17}$$

and has a covariance matrix

$$v(\hat{w}_a) = \sigma^2 (X_a' \tilde{\Sigma}^{-1} X_a)^{-1}. \tag{4.18}$$

The matrix $\tilde{\Sigma}$ is the covariance matrix of $\tilde{y} = C Z_J + \tilde{v}$, and as CZ_J is a constant, this matrix is also the covariance matrix of $\tilde{v}$. From (2.9)

$$\tilde{v} = T_2^{-1} \tilde{u} \tag{4.19}$$

and therefore

$$\tilde{\Sigma} = (T_2^{-1}) \tilde{\Omega} (T_2^{-1})' \tag{4.20}$$

Equation (4.20) allows to express (4.17) and (4.18) using $\tilde{\Omega}$, as

$$\hat{w}_a = (X_a^{*'} \tilde{\Omega}^{-1} X_a^*)^{-1} (X_a^{*'} \tilde{\Omega}^{-1} u_C) \tag{4.21}$$

$$v(\hat{w}_a) = \sigma^2 (X_a^{*'} \tilde{\Omega}^{-1} X_a^*)^{-1} \tag{4.22}$$

where

$$X_a^* = T_2 X_a \tag{4.23}$$

is the result of applying the non-stationary operator T_2 to the columns of X_a, and

$$u_C = T_2 (y_C - CZ_J) = T_2 y_C - T_2 CZ_J \tag{4.24}$$

is the result of differencing the corrected series y_C and the correction term C. Then, the likelihood function will be

$$l(\theta | y) = (2\pi\sigma^2)^{-(\frac{m-k}{2})} |\tilde{\Omega}|^{-1/2} |X_a^{*'} \tilde{\Omega}^{-1} X_a^*|^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} (u_C - X_a^* \hat{w}_a)' \tilde{\Omega}^{-1} (u_C - X_a^* \hat{w}_a) \right\} \tag{4.25}$$

It is interesting to point out the relationship between the likelihoods in the stationary and non-stationary case. Comparing (4.12) to (4.25) we see that in both cases the likelihood

uses the covariance matrix for the complete stationary process, $\tilde{\Omega}$, and in both a correction term appears over the standard likelihood for the additive outlier model. (See Peña, 1990). In the stationary case this term is $|X_a'\tilde{\Omega}X_a|$, whereas in the nonstationary case is $|X_a^{*\prime}\tilde{\Omega}^{-1}X_a^*|$ where X_a^* is the result of applying the non stationary operators to the columns of X_a.

It is useful to concentrate the parameters Z_J out of the likelihood function in the same way as it has been done for the parameters w_a which estimate the missing values. Calling $\bar{Z}_*$ the vector obtained by introducing arbitrary values $\bar{Z}_J$ in the position corresponding to the missing Z_J, we have

$$\tilde{Z} = \begin{bmatrix} Z_* \\ Z_R \end{bmatrix} = \begin{bmatrix} \bar{Z}_* \\ \bar{Z}_R \end{bmatrix} - \begin{bmatrix} X_J & 0 \\ 0 & X_a \end{bmatrix} \begin{bmatrix} w_J \\ w_a \end{bmatrix} = \bar{Z}_T - X_T w_T \tag{4.26}$$

where $\bar{Z}_T$ is the corrected series by filling all the holes, and X_T is the matrix of all the dummy variables. Then

$$\tilde{v} = \tilde{Z}_R - AZ_* = \bar{Z}_R - X_a w_a - A(\bar{Z}_* - X_J w_J), \tag{4.27}$$

and using again (4.26)

$$\tilde{v} = [-A\ I]\,(\bar{Z}_T - X_T w_T) \tag{4.28}$$

where [-A I] is the $n \times n$ matrix obtained attaching an identity matrix of dimensions (n-d) to the matrix A. Also, by (4.20)

$$v'\Sigma^{-1}v = (\bar{Z}_T - X_T w_T)' \begin{bmatrix} T_1' \\ T_2' \end{bmatrix} \Omega^{-1} [T_1 T_2]\,(\bar{Z}_T - X_T w_T) \tag{4.29}$$

and using that $T_1 Z_* + T_2 Z_R = u$, for (2.6)

$$v'\Sigma^{-1}v = (u_C^* - X_T^* w_T)'\Omega^{-1}(u_C^* - X_T^* w_T) \tag{4.30}$$

where $u_C^* = T_1\bar{Z}_* + T_2\bar{Z}_R$ is the differenced series, and $X_T^* = [T_1X_J, T_2X_a]$ is the matrix obtained by differencing the columns of the X_T matrix. From (4.30) and (4.15), and after some straightforward algebra

$$\hat{w}_T = (X_T^{*\prime}\Omega^{-1}X_T^*)^{-1}(X_T^{*\prime}\Omega^{-1}u_C^*) \tag{4.31}$$

Therefore $\hat{w}_T = (\hat{w}'_J, \hat{w}'_a)'$ can be concentrated out of the likelihood (4.25) and

$$l(\theta|y) = (2\pi\sigma^2)^{\frac{m-k}{2}} |\Omega|^{-1/2} |X_a^{*\prime}\Omega^{-1}X_a^*|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(u_C^* - X_T^*\hat{w}_T)'\Omega^{-1}(u_C^* - X_T^*\hat{w}_T)\right\} \tag{4.32}$$

Note that in this concentrated likelihood the correction term only involves the missing observations for $t > d$, whereas in the exponent the entire X_T^* matrix appears.

5. PERFORMANCE OF THE PROCEDURES

We have run a simulation experiment to compare both methods. To avoid differences in performance due to differences in the maximization routines or in the computation of the likelihood, the same algorithm have been used, when possible, in both methods. For instance, the likelihood is computed always using Melard (1984) algorithm, but in the first case (KL from now on) it is applied to the series, whereas in the second (PLM from now on) it is also applied to columns of the matrix X_T to concentrate the parameter w out of the likelihood. Then, in the first case (KL) the interpolation is carried out by using the fixed interval algorithm, whereas in the second (PLM) the interpolation is obtained by solving the regression linear equations with the QR algorithm.

In order to check the loss of precision when dropping the correction term in the likelihood in the PLM method, we have also included under the heading AI (Intervention Analysis) the results for this case. The name is, of course, because then the likelihood used is the same as in the standard Intervention Analysis model. All the computation has been done with a 486 PC, and the length of the series simulated has always been 100.

M O D E L		one missing (at 50)			five missing (41 to 45)		
		$\bar{e}$	MSE	t	$\bar{e}^{(*)}$	$MSE^{(*)}$	t
AR(1)	KF	-.0029	.5883	317	-.02284	1.328	326
$\phi = .8$	PLM	-.0029	.5883	344	-.02285	1.328	798
	AI	-.0035	.5883	343	-.02314	1.328	800
MA(1)	KF	-.0027	.5266	382	.0078	1.297	394
$\theta = .7$	PLM	-.0097	.5804	420	.0078	1.297	965
	AI	-.0001	.5303	436	.0077	1.299	1023
ARIMA	KF	-.0081	.2016	363	-.0191	1.337	376
(1,1,0)	PLM	-.0081	.2016	348	-.0191	1.337	781
$\phi = .8$	AI	-.0081	.2016	348	-.0190	1.337	781

Table 1

Results of the simulation experiment. $\bar{e}$ is the mean error and $\bar{e}^{(*)}$ the mean of the five mean errors for the five missing. With the same notation, MSE is the mean square error and $MSE^{(*)}$ the mean of the five mean square errors. t is the time in seconds elapsed in the 1000 simulations.

Table 1 presents the mean, variance and square error of the interpolation error for 1000 simulations of the three methods considered (KF, PLM, IM), with three different models and two structures of missing data. It can be seen that the accuracy of the three methods is roughly the same, and, therefore, we can conclude that the correction term in the likelihood has a very small effect on the computations. The table indicates the total time required to carry out the 1000 simulations, the estimation of the parameters and the interpolation. It is clear that when the number of missing values is large the first procedure is the fastest. However, for a small number of outliers and a nonstationary model the PLM can be faster than the standard KL algorithm, as shown in the case of an ARIMA model with

a single missing value. The reason is that with a complete series we can use a very fast routine, as Melard (1984), to compute the likelihood, whereas if there are holes in the series the recursive routine is slower. This difference will be important for series with a large state space vector, as, for instance, monthly nonstationary seasonal data. On the other hand, when the number of outlier is very large this possible advantage will disappear because we need to apply the recursive routine to all the columns of the X matrix.

A conclusion from table 1 is that both procedures are very fast. For instance, to estimate the parameters and to interpolate five values in an ARIMA (1,1,0) model takes an average of 0,4 seconds with KF and 0,8 with PLM in a 486 PC machine. It is clear that when variations of speed in this range are not important, other factors should be consider.

The main advantage of the Peña-Ljung-Maravall procedure is its flexibility: (1) it allows to compute the covariance matrix of the interpolators directly, before doing any computations; (2) it can be implemented easily in the version AI in many existing software; (3) it provides compact formulas for the estimators and, thus, leads to a deeper understanding of how the computations has been carried out.

6. REFERENCES

[1] Anderson, B.D.A. and Moore, J.B. (1979). *Optimal Filtering*, Prentice-Hall.

[2] Ansley, C.F. and Kohn, R. (1983). 'Exact likelihood of vector autoregressive-moving average process with missing or aggregated data,' *Biometrika*, 70, 275-8.

[3] Bell, W. (1984). 'Signal Extraction for Nonstationary Series,' *The Annals of Statistics*, 12, 646-664.

[4] Bell, W. and Hillmer, S. (1991). 'Initializing the Kalman Filter for Nonstationary Time Series Models,' *Journal of Time Series Analysis*, 4, 283-300.

[5] Box, G.E.P. and Jenkins, G.F. (1976). *Time Series Analysis Forecasting and Control.* Holden day.

[6] Box, G.E.P. and Tiao, G.C. (1975). 'Intervention Analysis with applications to economic and environmental problems,' *Journal of the American Statistical Association*, 70, 70-79.

[7] Brubacher, S.R. and Tunnicliffe-Wilson, G. (1976). 'Interpolating time series with application to the estimation of holiday effects on electricity demand,' *Applied Statistics*, 25, 2, 107-116.

[8] Gómez, V. and Maravall, A. (1992). 'Estimation, Prediction and Interpolation for Nonstationary Series with the Kalman Filter,' Submitted for publication.

[9] Harvey, A.C. and Pierse, R.G. (1984). 'Estimating missing observations in economic time series,' *Journal of the American Statistical Association*, 79, 125-132.

[10] Jones, R.H. (1980). 'Maximum likelihood fitting of ARMA models to time series with missing observations,' *Technometrics*, 22, 3, 389-395.

[11] Kohn, R. and Ansley, C.F. (1983). 'Fixed interval estimation in state space models when some of the data are missing or aggregated,' *Biometrika*, 70,3, 683-8.

[12] Khon, R and Ansley, C.F. (1986). 'Estimation, Prediction, and Interpolation for ARIMA Models with Missing Data,' *Journal of the American Statistical Association*, 81, 751-761.

[13] Ljung, G.M. (1982). 'The likelihood function for a stationary Gaussian autoregressive-moving average process with missing observations,' *Biometrika*, 69, 1, 265-8.

[14] Ljung, G.M. (1989). 'A note on the estimation of missing values in time series,' *Communication in Statistics, Simulation and Computation*, 18, 2, 459-465.

[15] Miller, R.B. and Ferreiro, O.M. (1984). 'A strategy to complete a time series with missing observations,' *Lectures Notes in Statistics*, 25, 251-275. Springer-Verlag, New-York.

[16] Melard, G. (1984). 'A Fast Algorithm for the exact likelihood of Autoregressive-Moving Average Models,' *Applied Statistics*, 35, 104-114.

[17] Peña, D. (1987). 'Measuring the importance of outliers in ARIMA models,' in *New Perspectives in Theoretical and Applied Statistics*, eds. M.L. Puri et al, Wiley, 109 - 118.

[18] Peña, D. (1990). 'Influential observations in time series,' *Journal of Business and Economic Statistics, 8*, 2, 235-241.

[19] Peña, D. and Tiao, G.C. (1991). 'A Note on Likelihood Estimation of Missing Values in Time Series,' *The American Statistician*, 45, 212-213.

[20] Peña, D. and Maravall, A. (1991). 'Interpolations, Outliers and Inverse Autocorrelations,' *Communications in Statistics (Theory and Methods)*, 20, 3175-3186.

[21] Pourahmadi, M. (1989). 'Estimation and interpolation of missing values of a stationary time series,' *Journal of Time Series Analysis*, 10, 2, 149-169.

[22] Wincek, M.A. and Reinsel, G.C. (1986). 'An exact maximum likelihood estimation procedure for regression-ARMA time series models with possibly nonconsecutive data,' *Journal of the Royal Statistical Society*, 48, 3, 303-313.

ACKNOWLEDGEMENTS

Daniel Peña acknowledges support from DGICYT, Spain, project PB90-0266.